
Differentiable Probabilistic Models

William Watson
nextbillyonair@gmail.com

Abstract

d

Contents

1	Introduction	7
1.1	Philosophy	7
2	Preliminary	7
2.1	Kronecker Product	7
2.2	Gradients	7
2.3	Jacobian	7
2.4	Hessian	7
2.5	Newton Optimization	7
3	Distributions	7
3.1	Distribution	7
3.2	Arcsine	7
3.3	Bernoulli	7
3.4	Beta	7
3.5	Categorical	7
3.6	Cauchy	7
3.7	Chi Square	7
3.8	Conditional Model	7
3.9	Convolution	7
3.10	Data	7
3.11	Dirac Delta	7
3.12	Dirichlet	7
3.13	Exponential	7
3.14	Fisher-Snedcor (F-Distribution)	7
3.15	Gamma	7
3.16	Generator	7
3.17	Gumbel Softmax	7
3.18	Gumbel	7
3.19	Half Cauchy	7
3.20	Half Normal	7
3.21	Hyperbolic Secant	7
3.22	Langevin	7
3.23	Laplace	7
3.24	Log Cauchy	7
3.25	Log Laplace	7
3.26	Log Normal	7
3.27	Logistic	7
3.28	Normal (Multivariate)	7

3.29	Rayleigh	7
3.30	Relaxed Bernoulli	7
3.31	Student T	7
3.32	Transformed Distribution	7
3.33	Uniform	7
4	Mixture Models	7
4.1	Mixture Model	8
4.2	Gumbel Mixture Model	8
4.3	Infinite Mixture Model	8
5	Transforms	8
5.1	Transform	8
5.2	Inverse Transform	8
5.3	Affine	8
5.4	Exp	8
5.5	Expm1	8
5.6	Gumbel	9
5.7	Log	9
5.8	Logit	9
5.9	Power	9
5.10	Reciprocal	10
5.11	Sigmoid	10
5.12	SinhArcsinh	10
5.13	Softplus	10
5.14	Softsign	10
5.15	Square	10
5.16	Tanh	10
6	Criterion and Divergences	10
6.1	Cross-Entropy	10
6.2	Perplexity	10
6.3	Forward KL Divergence	11
6.4	Reverse KL Divergence	11
6.5	Jensen-Shannon Divergence	11
6.6	Earth Mover's Distance	11
7	ELBO	11
8	Adversarial Loss	11
8.1	Adversarial Loss	13

8.2	GAN Loss	13
8.3	MMGAN Loss	13
8.4	WGAN Loss	13
8.5	LSGAN Loss	13
8.6	Gradient Penalty	13
8.7	Spectral Norm	13
9	Models	13
9.1	Base Models	13
9.1.1	Model	13
9.2	Regression	13
9.2.1	Linear Regression (Normal)	13
9.2.2	L1 Regression (Laplace)	13
9.2.3	Ridge Regression (Normal + Normal Prior on Weights)	13
9.2.4	Lasso Regression (Normal + Laplace Prior on Weights)	13
9.3	Classification	13
9.3.1	Logistic Regression (Bernoulli)	13
9.3.2	Bayesian Logistic Regression (Bernoulli)	13
9.3.3	Softmax Regression (Categorical)	13
9.4	Clustering	13
9.4.1	Gaussian Mixture	13
9.5	Probabilistic Matrix Factorization	13
9.6	Generative Adversarial Networks	13
9.6.1	Generative Adversarial Networks	13
9.6.2	GAN Model	13
9.6.3	MMGAN Model	13
9.6.4	WGAN Model	13
9.6.5	LSGAN Model	13
10	Monte Carlo	13
10.1	Monte Carlo Approximation	13
10.2	Linear Congruential Generator	13
10.3	Inverse Transform Sampling	13
10.4	Box-Muller	13
10.5	Marsaglia-Bray	13
10.6	Rejection Sampling	13
11	Markov Chain Monte Carlo (MCMC)	13
11.1	Metropolis	13
11.2	Metropolis-Hastings	13
11.3	Metropolis-Adjusted Langevin Algorithm (MALA)	13

11.4 Hamiltonian Monte Carlo	13
--	----

1 Introduction

1.1 Philosophy

2 Preliminary

2.1 Kronecker Product

2.2 Gradients

2.3 Jacobian

2.4 Hessian

2.5 Newton Optimization

3 Distributions

3.1 Distribution

3.2 Arcsine

3.3 Bernoulli

3.4 Beta

3.5 Categorical

3.6 Cauchy

3.7 Chi Square

3.8 Conditional Model

3.9 Convolution

3.10 Data

3.11 Dirac Delta

3.12 Dirichlet

3.13 Exponential

3.14 Fisher-Snedcor (F-Distribution)

3.15 Gamma

3.16 Generator

3.17 Gumbel Softmax

3.18 Gumbel

3.19 Half Cauchy

3.20 Half Normal

3.21 Hyperbolic Secant

3.22 Langevin

3.23 Laplace

3.24 Log Cauchy

3.25 Log Laplace

3.26 Log Normal

3.27 Logistic

4.1 Mixture Model

4.2 Gumbel Mixture Model

4.3 Infinite Mixture Model

5 Transforms

Transforms are invertible functions that can be applied to a random variable to change the distribution.

5.1 Transform

5.2 Inverse Transform

5.3 Affine

- **Parameters**

- Location $\mu \in \mathbb{R}^n$
- Scale $\sigma > 0$

- **Forward**

$$f(x) = \mu + \sigma \cdot x \quad (1)$$

- **Inverse**

$$f^{-1}(y) = \frac{y - \mu}{\sigma} \quad (2)$$

- **Log Absolute Determinant Jacobian**

$$\log |\det \mathbf{J}|(x, y) = \log |\sigma| \quad (3)$$

5.4 Exp

- **Parameters**

- None

- **Forward**

$$f(x) = e^x \quad (4)$$

- **Inverse**

$$f^{-1}(y) = \log y \quad (5)$$

- **Log Absolute Determinant Jacobian**

$$\log |\det \mathbf{J}|(x, y) = x \quad (6)$$

5.5 Expm1

- **Parameters**

- None

- **Forward**

$$f(x) = e^x - 1 \quad (7)$$

- **Inverse**

$$f^{-1}(y) = \log(1 + y) \quad (8)$$

- **Log Absolute Determinant Jacobian**

$$\log |\det \mathbf{J}|(x, y) = x \quad (9)$$

5.6 Gumbel

- **Parameters**

- Location $\mu \in \mathbb{R}^n$
- Scale $\sigma > 0$

- **Forward**

$$f(x) = \exp \left(- \exp \left(- \frac{x - \mu}{\sigma} \right) \right) \quad (10)$$

- **Inverse**

$$f^{-1}(y) = \mu - \sigma \cdot \log(-\log(y)) \quad (11)$$

- **Log Absolute Determinant Jacobian**

$$\log |\det \mathbf{J}|(x, y) = -\log \left(\frac{\sigma}{-\log(y) \cdot y} \right) \quad (12)$$

5.7 Log

- **Parameters**

- None

- **Forward**

$$f(x) = \log x \quad (13)$$

- **Inverse**

$$f^{-1}(y) = \exp y \quad (14)$$

- **Log Absolute Determinant Jacobian**

$$\log |\det \mathbf{J}|(x, y) = -y \quad (15)$$

5.8 Logit

- **Parameters**

- None

- **Forward**

$$f(x) = \log \left(\frac{x}{1-x} \right) \quad (16)$$

- **Inverse**

$$f^{-1}(y) = \frac{1}{1 + e^{-y}} \quad (17)$$

- **Log Absolute Determinant Jacobian**

$$\log |\det \mathbf{J}|(x, y) = \log(1 + e^{-y}) + \log(1 + e^y) \quad (18)$$

5.9 Power

- **Parameters**

- Power p

- **Forward**

$$f(x) = \begin{cases} e^x & p = 0 \\ (1 + x \cdot p)^{1/p} & \text{otherwise} \end{cases} \quad (19)$$

- **Inverse**

$$f^{-1}(y) = \begin{cases} \log y & p = 0 \\ y^{p-1}/p & \text{otherwise} \end{cases} \quad (20)$$

- **Log Absolute Determinant Jacobian**

$$\log |\det \mathbf{J}|(x, y) = \begin{cases} x & p = 0 \\ \left(\frac{1}{p} - 1 \right) \cdot \log(x \cdot p + 1) & \text{otherwise} \end{cases} \quad (21)$$

5.10 Reciprocal

- **Parameters**

- None

- **Forward**

$$f(x) = 1/x \quad (22)$$

- **Inverse**

$$f^{-1}(y) = 1/y \quad (23)$$

- **Log Absolute Determinant Jacobian**

$$\log |\det \mathbf{J}|(x, y) = -2 \cdot \log |x| \quad (24)$$

5.11 Sigmoid

- **Parameters**

- None

- **Forward**

$$f(x) = \frac{1}{1 + e^{-x}} \quad (25)$$

- **Inverse**

$$f^{-1}(y) = \log \left(\frac{y}{1 - y} \right) \quad (26)$$

- **Log Absolute Determinant Jacobian**

$$\log |\det \mathbf{J}|(x, y) = -\log(1 + e^{-x}) - \log(1 + e^x) \quad (27)$$

5.12 SinhArcsinh

5.13 Softplus

5.14 Softsign

5.15 Square

5.16 Tanh

6 Criterion and Divergences

The criterion and divergences listed here can be used to quantify the "distance" between two distributions. Hence, in conjunction with torch optimizers, one can minimize said difference to learn the parameters of a distribution. For sake of notation clarity, p is the true distribution and q is the learned distribution. Hence we "fit" q to match p . In addition, we provide the Monte Carlo approximation.

6.1 Cross-Entropy

$$\begin{aligned} H(p, q) &= - \int p(x) \log q(x) dx \\ &= - \frac{1}{n} \sum_{x \sim p} \log q(x) \end{aligned} \quad (28)$$

6.2 Perplexity

$$\begin{aligned} H(p, q) &= \exp \left(- \int p(x) \log q(x) dx \right) \\ &= \exp \left(- \frac{1}{n} \sum_{x \sim p} \log q(x) \right) \end{aligned} \quad (29)$$

6.3 Forward KL Divergence

$$\begin{aligned} H(p, q) &= \int p(x) \log \frac{p(x)}{q(x)} dx \\ &= \frac{1}{n} \sum_{x \sim p} \log \frac{p(x)}{q(x)} \end{aligned} \tag{30}$$

6.4 Reverse KL Divergence

$$\begin{aligned} H(p, q) &= \int q(x) \log \frac{q(x)}{p(x)} dx \\ &= \frac{1}{n} \sum_{x \sim q} \log \frac{q(x)}{p(x)} \end{aligned} \tag{31}$$

6.5 Jensen-Shannon Divergence

6.6 Earth Mover's Distance

7 ELBO

8 Adversarial Loss

Adversarial Losses are criterion functions that allow for sample-sample based training between models p and q . More formally, it hides a Discriminator model that attempts to discriminate between the real data from p and fake data generated from q .

8.1 Adversarial Loss

8.2 GAN Loss

8.3 MMGAN Loss

8.4 WGAN Loss

8.5 LSGAN Loss

8.6 Gradient Penalty

8.7 Spectral Norm

9 Models

9.1 Base Models

9.1.1 Model

9.2 Regression

9.2.1 Linear Regression (Normal)

9.2.2 L1 Regression (Laplace)

9.2.3 Ridge Regression (Normal + Normal Prior on Weights)

9.2.4 Lasso Regression (Normal + Laplace Prior on Weights)

9.3 Classification

9.3.1 Logistic Regression (Bernoulli)

9.3.2 Bayesian Logistic Regression (Bernoulli)

9.3.3 Softmax Regression (Categorical)

9.4 Clustering

9.4.1 Gaussian Mixture

9.5 Probabilistic Matrix Factorization

9.6 Generative Adversarial Networks

9.6.1 Generative Adversarial Networks

9.6.2 GAN Model

9.6.3 MMGAN Model

9.6.4 WGAN Model

9.6.5 LSGAN Model

10 Monte Carlo

10.1 Monte Carlo Approximation

10.2 Linear Congruential Generator

10.3 Inverse Transform Sampling

10.4 Box-Muller

10.5 Marsaglia-Bray

10.6 Rejection Sampling

11 Markov Chain Monte Carlo (MCMC)